Scalable Question Generation for Evaluating Longitudinal Reasoning in Electronic Health Records

Jordan L. Cahoon[1,2] CAHOON@STANFORD.EDU
Chloe O. Stanwyck[1,3] CHLOEO@STANFORD.EDU
Kevin Keet[4] KKEET@STANFORD.EDU
Sulaiman Somani [4] SSOMANI@STANFORD.EDU
Alison Callahan[4] ACALLAHA@STANFORD.EDU
Jason A. Fries[1,4] JFRIES@STANFORD.EDU
Nigam H. Shah[1,4,5] NIGAM@STANFORD.EDU
Emily Alsentzer[1,7] EMA2016@STANFORD.EDU
[1]*Department of Biomedical Data Science, Stanford School of Medicine, Stanford, CA, USA*
[2]*Department of Pathology, Stanford School of Medicine, Stanford, CA, USA*
[3]*Department of Anesthesiology, Perioperative and Pain Medicine, Stanford School of Medicine, Stanford, CA, USA*
[4]*Department of Medicine, Stanford, Stanford School of Medicine, CA, USA*
[5]*Center for Clinical Excellence Research, Stanford School of Medicine, Stanford, CA, USA*
[6]*Department of Computer Science, Stanford University, Stanford, CA, USA*

Abstract

Evaluating question-answer systems for electronic health records is challenging due to the high cost of annotation, limiting the realism and scale of existing benchmarks. In this work, we introduce a scalable large language model-generated, clinician-verified framework to automatically generate questions that evaluate information retrieval over longitudinal records. This framework leverages patient timelines to generate questions that emulate questions asked during chart review. We compare generation approaches that leverage a single History & Physical (H&P) note versus supplementing the H&P with patient facts. Physicians approved 93% of questions generated from the H&P with patient facts, a 7% increase from using the H&P alone. Incorporating facts into the generation process yielded a 4% increase in verifiable questions and a 30% increase in multi-hop questions, which are the most clinically useful questions that synthesize information across multiple encounters. Our findings demonstrate the utility of our framework to support meaningful evaluations of clinical question-answer system performance at scale.

Keywords: question generation, EHRs

Data and Code Availability This work leverages de-identified electronic health data from the STAnford Research Repository (STARR). The code used to implement our methods is available here.

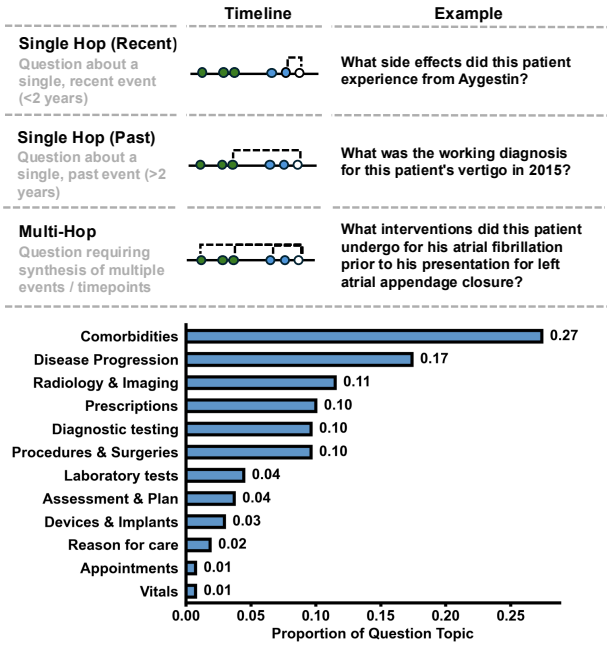

Figure 1: Generated question categories.

Institutional Review Board (IRB) Data access and chart review were conducted under IRB-57916 and IRB-78065 respectively.

1. Introduction

Large language models (LLMs) are rapidly transforming clinical workflows throughout hospitals. At

a growing number of institutions, clinicians can now send requests about specific patient medical records to secure LLM deployments embedded in electronic health record (EHR) systems (Armitage, 2025; Diaz, 2025; Lynn, 2025). Through these interactions, clinicians may ask questions about a patient's medical history, generate summaries, or retrieve specific information from far back in the chart. Such use cases require models to navigate and synthesize information spread across long and complex patient records. Yet, despite advances in long context modeling, it remains unclear how well LLMs can reason over this extended text, particularly in the clinical domain.

Robust evaluation of LLM clinical question-answer (QA) systems requires identifying a diverse set of realistic questions and extracting gold standard answers directly from patient records. Due to the information dense nature of EHRs, this is a time consuming process even for expert clinician annotators. The manual effort required to curate such datasets constrains their scale, which in turn limits the diversity of tasks represented in benchmarks.

Identifying realistic questions is the first step in developing a robust benchmark. Instead of solely relying on human annotation, we propose an LLM-based, clinician-verified approach to generate topically diverse, clinically useful, and temporally varied questions for evaluating clinical QA systems. We organize these questions into categories based on temporal framing and reasoning - single-hop (recent), single-hop (past), and multi-hop - and clinician-curated question topics (see Figure 1). This categorization enables downstream evaluation of clinical QA systems for different approaches like long-context modeling, retrieval augmented generation, and agentic frameworks. Concretely, we contribute the following:

- **Evaluation-Driven Questions**: Present a framework for generating, categorizing, and verifying realistic questions from patient encounters
- **Strong Approval**: Questions have 93% clinician approval rating. Incorporating patient timeline facts alongside History & Physical (H&P) notes yields a 4% increase in verifiable questions and a 30% increase in multi-hop questions compared to H&P notes alone.
- **Scalability**: Develop an LLM-as-a-Judge method with clinician agreement comparable to that within clinician annotators, enabling the large scale generation of high quality questions.

2. Related Work

2.1. Question Generation

Manually curating challenging questions that require multiple levels of reasoning can be challenging at scale [CITE THE MULTI HOP Q PAPERS X2]. Combined with the expanding context windows of modern LLMs, effective benchmark questions must be able to reason over lengthy texts which can be difficult even for expert annotators. Loong is a multi-document benchmark that uses free annotation and templates to generate question that require reasoning over long contexts (Wang et al., 2024). More recently, Gill et al. (2025) have detailed limitations in producing synthetic benchmarks by showing that synthetic benchmarks are easier for LLMs than the human curated counterparts. In our work, we adapt these question generation techniques to generate challenging, realistic questions that reason over longitudinal patient records.

Several datasets exist for benchmarking EHR QA tasks. EHRNoteQA is a LLM generated, clinician modified and verified benchmark built using MIMIC-IV EHR (Johnson et al., 2023; Kweon et al., 2024). While EHRNoteQA encompasses many question types, the questions only reason over a few discharge summaries which represents a small portion of a patient's complete medical timeline. MedAlign is a clinician generated QA dataset with longitudinal records, but only contains a small number of retrieval questions, lacking the diversity of questions necessary to construct a comprehensive benchmark (Fleming et al., 2024). Both EHRNoteQA and MedAlign focus on questions that are topically diverse and clinically useful, but lack a large set of questions that are *temporally varied*, synthesizing information over different parts of the chart. Most recently, TIMER-Bench introduced an LLM-generated, clinician-verified benchmark to evaluate models' abilities to reason over longitudinal records across different temporal spans (Cui et al., 2025). While TIMER-Bench focuses on distributing questions across the patient timeline to test temporal reasoning, it does not emphasize questions that a clinician might plausibly ask during a specific visit, limiting the utility of leveraging this approach for evaluating clinical QA systems. In contrast, our framework grounds question generation in specific patient encounters by using H&P notes and categorizes the questions to identify where QA systems may fail through realistic scenarios. This work will support the release of a publicly available benchmark.

3. Data

Our goal was to identify a cohort that represents a wide variety of clinical scenarios to ensure we generate topically diverse questions. To ground each question in a realistic clinical context, we select the time of the question in a specialty-agnostic manner. Specifically, question generation is tied to each patient's most recent History & Physical (H&P) note, which summarizes both past and present medical information at the time of the visit.

Using a de-identified EHR database from a large research hospital, we sampled 25 patients that had at least 100 notes prior to their most recent H&P note. This ensures our cohort only includes patients with sufficient information to generate complex questions that exceed the capacity of manual question generation. Notes were filtered before the reference H&P note timestamp to simulate the chart review process.

4. Methods

We introduce a scalable framework to generate and verify realistic clinical questions grounded by patient visits. The framework is organized into three modules: Fact, Question, and LLM-as-Judge. For all generative tasks in each module, we used a PHI-compliant instance of Gemini-Pro 2.5 Chat, which achieved stellar performance on the LMArena long-query and overall text leader boards (Chiang et al., 2024).

4.1. Patient Timeline Construction

Longitudinal clinical notes were used as the foundation for question generation. However, raw clinical notes are flawed documents, containing remnants of formatting, duplicated information due to copy-forward, clinical shorthand, and temporal ambiguity from interleaved information recorded in other visits. Altogether, LLMs struggle to reason over raw clinical notes. "Note bloat," where the note contains text that does not add new information, further limits the utility of using raw clinical notes by restricting how much information can fit in context.

We circumvented these issues by constructing patient timelines from raw clinical notes. A timeline is defined as a list of atomic claims, logical propositions that cannot be further decomposed, accompanied by timestamps (e.g. "Patient was diagnosed with Type 2 Diabetes (2021-09-13)") (Russell, 2009). Prior work has demonstrated that LLMs excel at extracting atomic claims from clinical notes, which in turn can be interpreted as patient *facts* (Chung et al., 2025; Munnangi et al., 2024). Distilling notes into a list of facts removes formatting and clinical shorthand from the raw note without changing the overall information.

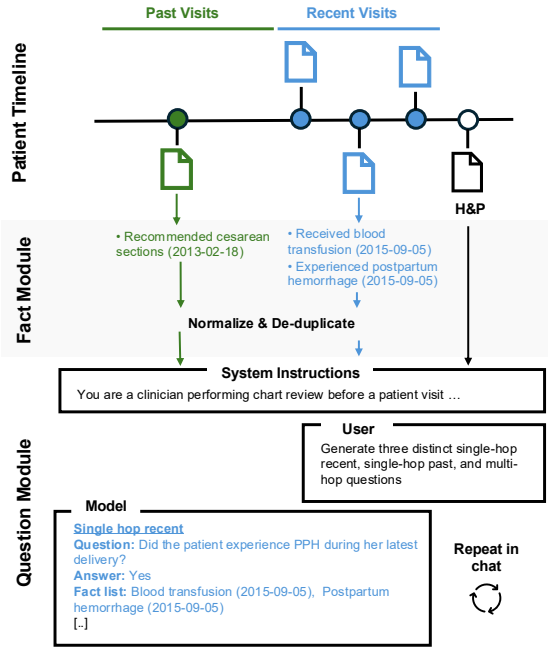

Figure 2: Question Generation framework.

The Fact Module extracted facts with timestamps from notes prior to the most recent H&P note (Appendix A, Figure 2). The fact timestamps were normalized to the estimate date of the event and the resulting fact list was de-duplicated to remove facts that refer to the same event (Appendix A). We observe a 3 times reduction in the average amount of tokens for refined facts from raw notes (See Table 2).

4.2. Question Generation

The Question Module generates free-text QA pairs designed to extract information across different temporal scopes, ordered by increasing difficulty: single-hop (recent), single-hop (past), and multi-hop. Single-hop questions surface information from a single visit. This category was further divided to account for the recency bias observed in clinically relevant questions. During chart review, useful information is often concentrated closer to the time of the visit, hence we define single-hop (past) questions as those surfacing a single event that occurred more than two years prior to the reference H&P note timestamp.

User messages for question generation were structured as sequential chat dialogues. In each message,

the model was prompted to simultaneously generate queries for different question types ("labels"), encouraging the production of temporally varied questions (Raman et al., 2024). Previous user–model exchanges were appended to subsequent prompts to further diversify the topical space. The Question Module also identified up to three relevant clinical topics for each question that were later verified by clinicians (see Figure 2, Table 1).

We evaluated two input strategies for grounding generation in clinical context: (1) a single-note approach using the H&P note alone, and (2) a patient timeline approach incorporating refined facts (Subsection 4.1). We compare these two methods to assess commonly used approaches from previous clinical QA generation methods (Section 2). The Question Module also generates the answer and the supporting facts required to derive it, either from the refined facts or the H&P note.

4.3. Question Verification

Three board-certified physicians assessed the quality of generated questions. When annotating the questions, clinicians had access to the medical record number (MRN, patient identifier) and other relevant visit information (Appendix C). The MRN allowed clinicians to view the complete patient chart when conducting their annotations to ensure the question was specific to that patient, despite the data being de-identified. Importantly, the clinicians were not made aware of the two generation approaches (H&P and H&P with facts).

Clinicians assessed the question quality through a series of binary assessments and ensured the question topic and reasoning was correctly categorized. We define a high-quality question as one that is clinically relevant to the specific patient, verifiable, free of leakage (i.e., excluding visit-specific details from the H&P note that would not be accessible during review), and consistent with the patient chart. To assess annotator agreement, 54 questions were annotated by two clinicians (Appendix C).

In addition to manual annotation, the LLM-as-a-Judge Module assessed if the question was patient specific, had a verifiable answer, and had natural phrasing. To help assess clinical relevance, we also included the complete H&P Note. See Appendix D for more details.

5. Results

Clinicians reviewed 108 questions, divided between the two generation approaches. Based on the evaluation criteria in Subsection 4.3, 93% and 86% of questions generated with the H&P with facts and H&P only approaches, respectively, were approved for evaluating clinical QA systems (Table 1). Questions consistently achieved high marks for being patient-specific, verifiable, and consistent. We also observe that generated questions were topically diverse with a Simpson's diversity index, the probability of randomly drawing two questions with different gold-standard topics, of 0.95. Overall, both methods struggled with creating questions with natural phrasing and correctly identifying the clinical topics asked by the questions.

When examining the types of questions that clinicians did not approve for evaluation, we observe a difference between the generation approaches. Of the rejected questions, 63% were generated with only the H&P. Notably, half of these rejected questions were multi-hop, a format we consider the most clinically useful. In contrast, only 16% of rejected questions from the fact-supplemented approach were multi-hop. Taken together, these results show that incorporating longitudinal data improves realistic, complex question generation.

Annotator agreement was assessed on 54 questions evaluated by two clinicians. Clinicians provided the same approval response for 81% of questions, and for every question, there was at least one metric on which both clinicians agreed. The average Cohen's $\bar{\kappa}$ across all assessment categories was 0.32 (Table 3), consistent with fair agreement for complex clinical annotation tasks. Agreement on Overall Approval was lower ($\bar{\kappa} = 0.04$, Figure 14), reflecting the inherent subjectivity of determining which questions are most clinically valuable to ask at a visit. Importantly, when approval was defined more explicitly—as requiring questions to be patient-specific, verifiable, consistent, and free of leakage—agreement rose to $\bar{\kappa} = 0.38$ (Figure 15). We view these differences as expected: clinicians from different specialties may value different types of questions, and a question only needs to be useful to one clinician to be meaningful in practice. This underscores both the diversity of clinical reasoning and the importance of including a broad range of questions in evaluation datasets.

We also examined annotator agreement between clinicians and the LLM-as-a-Judge module. Across all metrics, the LLM-as-a-Judge achieved annotator agreement that was at least as high as the agreement observed between pairs of clinicians. Detailed results are provided in Appendix D. These findings suggest

Table 1: Clinician evaluation of generated questions

	Assessment	H&P+Facts	H&P
Quality	Patient specific	98	98
	Verifiable answer	**98**	94
	Consistent with chart	99	**100**
	Leakage with H&P	/	6
	Natural phrasing	**61**	50
Type	Correct temporality	**93**	89
	Correct topics	**78**	75
	Overall approval	**93**	86

that using an LLM-as-a-Judge method may support manual evaluation. In practice, this capability may enable LLMs to filter out low-quality questions, scaling the generation of high quality questions for benchmarking.

6. Discussion & Future Work

We present an LLM-generated, clinician-verified framework for producing questions that are topically diverse, clinically useful, and temporally varied. To ensure quality, we combined clinician review with LLM-as-a-Judge verification. For scalability, question generation was grounded in H&P notes, mirroring the way clinicians review charts during patient visits. This framework is flexible and can be adapted to other settings. Our current work focuses primarily on clinical notes, excluding information from structured data, such as labs and medications. A natural extension would incorporate structured elements to enrich question generation. Ultimately, clinically grounded questions are essential for robust evaluation; without them, benchmarks cannot meaningfully assess readiness for deployment. In the future, we plan to validate model-generated answers and supporting facts. The resulting dataset will be released as a public benchmark to support the development and evaluation of clinical QA systems.

References

Hanae Armitage. Clinicians can 'chat' with medical records through new AI software, ChatEHR, June 2025. URL https://med.stanford.edu/news/all-news/2025/06/chatehr.html. Section: Artificial Intelligence (AI).

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference, March 2024. URL http://arxiv.org/abs/2403.04132. arXiv:2403.04132 [cs].

Philip Chung, Akshay Swaminathan, Alex J. Goodell, Yeasul Kim, S. Momsen Reincke, Lichy Han, Ben Deverett, Mohammad Amin Sadeghi, Abdel-Badih Ariss, Marc Ghanem, David Seong, Andrew A. Lee, Caitlin E. Coombes, Brad Bradshaw, Mahir A. Sufian, Hyo Jung Hong, Teresa P. Nguyen, Mohammad R. Rasouli, Komal Kamra, Mark A. Burbridge, James C. McAvoy, Roya Saffary, Stephen P. Ma, Dev Dash, James Xie, Ellen Y. Wang, Clifford A. Schmiesing, Nigam Shah, and Nima Aghaeepour. VeriFact: Verifying Facts in LLM-Generated Clinical Text with Electronic Health Records, January 2025. URL http://arxiv.org/abs/2501.16672. arXiv:2501.16672 [cs].

Hejie Cui, Alyssa Unell, Bowen Chen, Jason Alan Fries, Emily Alsentzer, Sanmi Koyejo, and Nigam Shah. TIMER: Temporal Instruction Modeling and Evaluation for Longitudinal Clinical Records, March 2025. URL http://arxiv.org/abs/2503.04176. arXiv:2503.04176 [cs].

Naomi Diaz. CHOP creates AI agent for Epic - Becker's Hospital Review | Healthcare News & Analysis, July 2025. URL https://www.beckershospitalreview.com/healthcare-information-technology/ehrs/chop-develops-ai-agent-for-epic/.

Scott L. Fleming, Alejandro Lozano, William J. Haberkorn, Jenelle A. Jindal, Eduardo Reis, Rahul Thapa, Louis Blankemeier, Julian Z. Genkins, Ethan Steinberg, Ashwin Nayak, Birju Patel, Chia-Chun Chiang, Alison Callahan, Zepeng Huo, Sergios Gatidis, Scott Adams, Oluseyi Fayanju, Shreya J. Shah, Thomas Savage, Ethan Goh, Akshay S. Chaudhari, Nima Aghaeepour, Christopher Sharp, Michael A. Pfeffer, Percy Liang, Jonathan H. Chen, Keith E. Morse, Emma P. Brunskill, Jason A. Fries, and Nigam H. Shah. MedAlign: A Clinician-Generated Dataset for Instruction Following with Electronic Medical Records. *Proceedings of the AAAI Conference on*

Artificial Intelligence, 38(20):22021–22030, March 2024. ISSN 2374-3468. doi: 10.1609/aaai.v38i20.30205. URL https://ojs.aaai.org/index.php/AAAI/article/view/30205. Number: 20.

Alexander Gill, Abhilasha Ravichander, and Ana Marasović. What Has Been Lost with Synthetic Evaluation?, June 2025. URL http://arxiv.org/abs/2505.22830. arXiv:2505.22830 [cs].

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, January 2023. ISSN 2052-4463. doi: 10.1038/s41597-022-01899-x. URL https://www.nature.com/articles/s41597-022-01899-x. Publisher: Nature Publishing Group.

Sunjun Kweon, Jiyoun Kim, Heeyoung Kwak, Dongchul Cha, Hangyul Yoon, Kwanghyun Kim, Jeewon Yang, Seunghyun Won, and Edward Choi. EHRNoteQA: An LLM Benchmark for Real-World Clinical Practice Using Discharge Summaries, November 2024. URL http://arxiv.org/abs/2402.16040. arXiv:2402.16040 [cs].

John Lynn. A Deep Dive Into the Announcements at Epic UGM 2025 | Healthcare IT Today, August 2025. URL https://www.healthcareittoday.com/2025/08/21/a-deep-dive-into-the-announcements-at-epic-ugm-2025/.

Monica Munnangi, Akshay Swaminathan, Jason Alan Fries, Jenelle Jindal, Sanjana Narayanan, Ivan Lopez, Lucia Tu, Philip Chung, Jesutofunmi A. Omiye, Mehr Kashyap, and Nigam Shah. FactEHR: A Dataset for Evaluating Factuality in Clinical Notes Using LLMs, December 2024. URL https://arxiv.org/abs/2412.12422v2.

Karthik Raman, Michael Bendersky, and Aditi Chaudhary. Its All Relative! – A Synthetic Query Generation Approach for Improving Zero-Shot Relevance Prediction. In *Findings of the Association for Computational Linguistics: NAACL 2024*, 2024. URL https://aclanthology.org/2024.findings-naacl.107.pdf.

Bertrand Russell. *The philosophy of logical atomism*. Routledge Classics. Routledge, Abingdon, Oxon, 2009. ISBN 978-0-415-47461-0.

Minzheng Wang, Longze Chen, Cheng Fu, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, Yunshui Li, Min Yang, Fei Huang, and Yongbin Li. Leave No Document Behind: Benchmarking Long-Context LLMs with Extended Multi-Doc QA, October 2024. URL http://arxiv.org/abs/2406.17419. arXiv:2406.17419 [cs].

Appendix A. Fact Module

The patient fact list was constructed in two phases: extraction and refinement. The Fact Module identified (1) the atomic claims and (2) the estimated timestamp that the claim occurred based on the context of the note. The real event timestamp needed to be estimated because clinical notes often reference events outside the documented visit. This step is important to reduce redundancies as multiple notes may describe the same event and precisely order facts in a patient timeline.

Facts were extracted directly from individual notes (See Figure 3). For very long notes, we chunked the note into 20k token chunks. While Gemini 2.5 Pro has a context window of 1 million tokens, we chunked notes to control for context length and enable parallelization. After fact extraction, the facts from all notes were merged in chronological order based on note sequence. On average, patients had fact lists of 3,300 entries (see Table 2.

We refined the fact list through an iterative process of batching and pruning to eliminate duplicate information (see Figure 4). The goal was to identify facts referring to the same information on the same date, thereby removing copy-forward content. The Fact Module removed redundant facts in batches of 500. To capture duplicates that might appear across widely separated time points, we shuffled and re-batched the list for three iterations. Overall this process removed roughly one-third of all extracted facts (see Table 2.

Fact extraction was limited to notes written within two years prior to the H&P timestamp. While this filtering may constrain the scope of generated questions, most patient-specific queries during chart review typically focus on recently documented information. By targeting this window, we aimed to generate clinically realistic questions while validating both clinician and LLM-as-a-Judge verification processes. Notably, generated questions still reason over information outside of the two year window, as the notes frequently reference past encounters. In future work, we plan to extend this approach to the full patient timeline, including structured data, which we expect will improve the quality and coverage of questions targeting older events.

Table 2: Count and number of tokens for each stage of fact extraction. The **Refined Facts**, pruned for redundancy, was used for question generation.

Stage	Count	Token Count
Raw Note	$(1.6 \pm 1.3) \times 10^3$	$(1.0 \pm 0.9) \times 10^5$
Fact	$(3.3 \pm 3.0) \times 10^3$	$(5.6 \pm 5.0) \times 10^4$
Refined Facts	$\mathbf{(2.1 \pm 1.9) \times 10^3}$	$\mathbf{(3.3 \pm 2.2) \times 10^4}$

Appendix B. Question Module

We conducted generation using two different types of inputs: H&P note and H&P note with facts derived from the patient timeline. The three question types-single hop (recent), single hop (past), and multi-hop-were generated concurrently. This processes was conducted twice for each patient, yielding in 12 questions per patient for the two input types. For the sampled 25 patients, this process generated 300 total questions. We share the abbreviated prompts used to generate questions in Figures 5,6. After generating the questions, we classified the clinical topics using a list verified by clinicians (Figures 7).

Appendix C. Clinical Annotation

Three physicians from anesthesiology, internal medicine, and cardiology were recruited to conduct annotations. Annotators were trained using a set of held out generated questions and discussed expected responses to improve agreement (See Figure 8 for annotation guidelines). During annotation, clinicians had access to the patient medical record number, reference H&P timestamp, question, answer, and supporting facts. Additionally, we provided visit information, including the visit type and LLM-extracted details from the H&P, the patient clinical summary and visit reason. Clinicians were aware that the answer, supporting facts, and generated visit information were not verified and were instructed to use the information with caution.

Each annotator assessed a total of 54 questions, 36 of which were assessed by another annotator. Annotators recorded their responses in Label Studio and used Epic Hyperspace to verify clinical accuracy and utility.

Fact Extraction Prompt

Task Definition
You are a clinician who is performing chart review on a patient who was just admitted to the hospital. Your task is to generate a list of atomic claims from the given excerpt of a clinical note.
Atomic Claim Definition
An atomic claim is a phrase or sentence that makes a single assertion. The assertion may be factual or may be a hypothesis posed by the text. Atomic claims are indivisible and cannot be decomposed into more fundamental claims. More complex facts and statements can be composed from atomic facts. Atomic claims should have a subject, object, and predicate. The predicate relates the subject to the object.
Do:

1. Extract discrete atomic claims from the `"text"` field. Each claim must include a subject, predicate, and object, and must stand alone without ambiguity.
2. Include only clinically relevant claims (symptoms, procedures, tests, medications, diagnoses, clinical locations).
3. Use only the provided text. Do not add outside knowledge or assumptions. Preserve the full context of each claim.
4. Write each claim in the shortest unambiguous form. Avoid pronouns or vague references.
5. Append a date (YYYY-MM-DD) to every claim:
   (a) If the text specifies an absolute date, use that date.
   (b) If the text uses a relative reference (e.g., "yesterday," "last week," "last month"), resolve it against the `note_date`.
   (c) If no event date is given, use the `note_date`.
   (d) For vague ranges (e.g., "last month"), default to the first day of that period unless the text specifies otherwise.
6. Always refer to the subject as "patient," even if the text uses the patient's name or identifiers.
7. If there are no valid clinically relevant claims, return `"claims"` as an empty list `[]`.

Do Not:

1. Do not include claims that are not directly about the patient's clinical care (e.g., provider names, note authors, addenda, phone numbers, clinic addresses, or administrative details).
2. Do not invent or infer claims beyond what is explicitly stated in the text.
3. Do not duplicate the `note_date` as the event date if the text already provides an event date.
4. Do not combine multiple events into a single claim — each claim must be atomic.
5. Do not include general background knowledge or medical facts not present in the text.

Format as a JSON
Input Schema: { "note_date": str, "text: str }
Output Schema: { "claims" : List[str] }
Examples Input: { "note_date" : "2021-01-15" "text": [...] }
Output: { "claims": ["The chief complaint documented was eye pain (2021-01-15)", "The patient reported that the left eye was red and looked like it was bleeding (2021-01-15)", "The patient used eye drops for the left eye (2021-01-15)", "Left eye redness partially cleared after using eye drops (2021-01-15)", "Left eye redness onset (2021-01-08)", [..]] }

Figure 3: Fact Extraction System Prompt. Prompt adjusted from Chung et al. (2025)

Fact De-duplication Prompt

Task Definition
You are an expert clinician tasked with reviewing a **list of patient facts**. Some of these facts may be duplicates or semantically redundant. Your job is to identify which facts should be removed so that the final fact list is concise, non-redundant, and still retains all unique clinical information.
You will not regenerate the fact list. Instead, you will **return the indices of facts to remove**.
—
Detailed Instructions
1. **Redundant fact definition**
* A fact is redundant if it asserts the same claim as another fact, even if phrased differently.
* Example: *“Patient has hypertension”* and *“History of high blood pressure”* → redundant.
* If two facts contain identical information except for timestamps, keep the most complete one (the one with more timestamps).
* If one fact is a subset of another (e.g., *“Admitted to hospital”* vs *“Admitted to hospital (2014-08-01)”*), mark the subset for removal.
2. **Conflicting facts**
* If two facts make contradictory claims, **do not** mark either as redundant. Both must be kept.
3. **Timestamps**
* Facts that differ only by **unique timestamps** are not redundant and must both be kept.
* Example: *“Admitted to hospital (2014-08-01)”* and *“Admitted to hospital (2014-09-01)”* → both should remain.
4. **Output format**
* Return a JSON object with one key:
“‘json ”redundant_fact_indices”: [¡list of 0-based indices to remove¿] “‘ * Do not return the facts themselves, only the indices. * If no redundancies are found, return an empty list:
“‘json ”redundant_fact_indices”: List[int] “‘
—
Example
Input facts (indexed for reference):
0: Patient has hypertension
1: History of high blood pressure
2: Admitted to hospital (2014-08-01)
3: Admitted to hospital (2014-09-01)
4: Admitted to hospital
Output
“‘json { “redundant_fact_indices”: [1, 4] } “‘
Example Explanation:

- Fact 1 is redundant with fact 0.
- Fact 4 is a subset of facts 2 and 3, so it is removed.
- Facts 2 and 3 are kept because they contain unique timestamps.

Do not include an explanation in your response.

Figure 4: Fact Deduplication System Prompt.

Question Generation System Prompt

You are a clinician who is seeing a new patient and you are performing a comprehensive review of the patient. Your job is to use the list of patient facts to generate retrieval-only questions, each with (a) a concise answer and (b) the exact supporting facts copied verbatim from the fact list.

Rules:

1. Use only the provided facts. Do not invent, summarize, interpret, calculate, or speculate.
2. Ensure temporal diversity across questions (cover early events, recent events, and multi-time point trends).
3. Avoid trivially simple questions; prefer those requiring reasoning or integration when appropriate.
4. Copy supporting facts exactly into fact_subset (no paraphrasing, trimming, or merging; include dates verbatim).
5. Phrase questions in natural clinical language, as a clinician would during admission or routine care. Avoid test-like wording and unnecessary precision.
6. Do not include exact dates unless required for verifiability. Prefer natural temporal references (e.g., "last 3 years," "most recent exam") or clinical context (e.g., "at admission").
7. Questions must allow an objective, unambiguous answer while remaining natural in clinical phrasing. Avoid vague prompts that lead to subjective answers (e.g., "details," "tell me about," "specifics"). Use anchored terms like "outcome," "result," or "findings" when appropriate.
8. Answers must be directly supported by the provided facts and must include all associated dates.
9. Ensure questions are clinically useful for admitting the patient (useful for decision-making, documentation, or patient understanding) and justify in the "clinical_relevance_rationale" field.
10. If no valid question can be generated, return an empty JSON array [].

Allowed topics include (non-exhaustive): Comorbidities; Procedures/Surgeries; [..]

Output format (strict): Return only a JSON array (no prose, no code fences).

Each object must follow:
{ "question_id": { "type": "integer" }, "reference_timestamp" : { "type": "string" }, "question_type": { "type": "string", "enum": ["single_hop_recent", "single_hop_past", "multi_hop"] }, "question": { "type": "string" }, "answer": { "type": "string" }, "fact_subset": { "type": "array", "items": { "type": "string" } }, "clinical_relevance_rationale": { "type": "string" } }

Question type definitions:

1. Single-hop recent: Answerable using a single event that occurred less than 2 years before the time of admission. [...]
2. Single-hop past: Answerable using a single event that occurred more than 2 years before the time of admission.
3. Multi-hop: Answerable only by combining information from two or more distinct events, typically from different time points relative to the time of admission. May involve linking causality, chronology, or comparison

Example 1: [..]

Figure 5: Example question generation prompt. Prompt was shortened for brevity. The in-context examples were generated by a board-certified clinician based on held-out patients.

Question Generation User Prompt

Generate three distinct complex retrieval questions that satisfy the constraints. The first must be "single-hop recent", the second "single-hop past", and the third "multi-hop". Do not generate questions about the patient's psychology, mental well being, or psychiatric care. Return only the JSON array in the specified schema and order: recent, past, multi-hop.

Figure 6: Example user generation prompt. Model responses and previous user prompts are appended for subsequent chat messages.

Metric	$\bar{\kappa}$
All categories	0.32
Patient Specific	0.67
Verifiable Answer	0.48
Consistent	0.67
Includes leakage	1.0
Natural Phrasing	0.13
Correct type	0.17
Correct category	0.18
Overall Approval	0.045
Overall Approval (Adjusted)	0.38

Table 3: Average Cohen'sκ values (2 sig figs) for each metric of assessment.

Overall annotators had a high approval rating for generated questions. While most metrics had fair agreement, others had lower annotator agreement (Table 3). For example, we observed a slight disagreement ($\bar{\kappa} = 0.045$) between annotators for overall approval into the benchmark (igure 14). The agreement improved ($\bar{\kappa} = 0.38$) when redefining approval questions for positive responses in Patient Specific, Verifiable Answers, Consistency and Contains no leakage, however there was still a slight discrepancy, likely due to the different training backgrounds of each physician 15. This highlights the need to include annotators from diverse specialties, as different clinicians may have different needs from clinical QA systems.

Appendix D. LLM-as-a-Judge Module

We developed an LLM-as-a-Judge Module to automatically assess the quality of the generated questions. Likewise with the clinicians, the LLM-as-a-Judge module had access to the reference H&P timestamp, visit related information (see Appendix C), answer, and supporting facts. Additionally, the module had access to the full H&P note. We note this is a limited view of the patient, as the module did not have access to the full chart, unlike the clinician annotators.

We selected three areas of assessment that could be validated without access to the complete patient chart: Patient Specific, Verifiable Question, and Natural Phrasing. We focused on these qualities because these were the biggest areas of improvement in initial testing. The prompts used for the LLM-as-a-Judge are shown in Figure 20. We ran the LLM-as-a-Judge module for 108 questions and analyzed how the responses compare to the three clinician annotators (Figure 16).

When examining responses for the Patient Specific and Verifiable Question assessments for questions with multiple annotations, we observe that the LLM-as-a-Judge modules responds the same with at least one of the annotators. The LLM-as-a-Judge framework had the highest agreement with Annotator A with $\bar{\kappa} = 1.0$ and $\bar{\kappa} = 0.32$ for Patient Specific and Verifiable Answer responses, respectively (Figures 17 and 18). On the other hand, we observe a low agreement with Annotator C, with $\bar{\kappa} = 0.0$ for both Patient Specific and Verifiable Answer responses (Figures 17 and 18). While this does not indicate strong agreement, this is the observed agreement of Annotator A and C, which suggests that the LLM-as-a-Judge Module represents a portion of clinician responses. In some cases, such as the Verifiable Question assessment, the LLM-as-a-Judge Module is a harsher critic for whether or not the question is verifiable. However, this comes with a tradeoff, as generated questions that include wording to improve verifiability (i.e. dates, times, locations) may not be naturally phrased. This confers with the responses of the Natural Phrasing assessment, where the LLM-as-a-Judge Module was more lenient than clinician

Question Topic Classification

You are a clinical information extraction system. Your task is to carefully analyze a patient-specific clinical question and identify the top 3 most relevant topics from a fixed list of categories.

Categories: Comorbidities, Procedures/Surgeries, Devices/Implants, Radiology/Imaging, Diagnostic testing (e.g. genetics, pathology, etc.), Demographics, Prescriptions (e.g. type, interactions, side effects, reasons not to medicate), Laboratory tests, Disease Progression status (e.g. severity, complications, staging information, functional status), Social Determinants of Health,Assessment & Plan, Vitals, Appointments, Family History Communications, Payment, Reason for care (e.g. reason for admission, referral etc.), Immunizations, Allergies, Other, None

Rules: 1. Select up to 3 topics that are most directly relevant to the question. 2. If fewer than 3 topics apply, fill the remaining slots with "None". 3. Always use the exact category names listed above. 4. Use "Other" sparingly. Try your best to use the given categories. 5. Use the answer to help inform topics of the question. Do not list topics in the answer.

Output must be in valid JSON format with the following structure:
Input format: { "question": str, "answer": str }
Output format: { "topic_1": str, "topic_2": str, "topic_3": str }
Example 1: Input: { "question": "What medications has this patient tried for her pelvic and rectal pain?", "answer": "Gabapentin, oral contraceptives (multiple names acceptable, incl. estrostep, blisovi FE, norethindrone, aygestin), nitroglycerin ointment" }
Output: { "topic_1": "Prescriptions", "topic_2": "Comorbidities", "topic_3": "None" }
Output must be in valid JSON format with the following structure:
Input format: { "question": str, "answer": str }
Output format: { "topic_1": str, "topic_2": str, "topic_3": str }

Example 1: Input: { "question": "What medications has this patient tried for her pelvic and rectal pain?", "answer": "Gabapentin, oral contraceptives (multiple names acceptable, incl. estrostep, blisovi FE, norethindrone, aygestin), nitroglycerin ointment" }
Output: { "topic_1": "Prescriptions", "topic_2": "Comorbidities", "topic_3": "None" }

Example 2: Input: { "question": "What airway-related issues has this patient experienced that should be noted prior to planning a general anesthetic?", "answer": "Congenital TEF and tracheal stenosis (repaired), esophageal anastomosis leak (resolved) and stricture, L vocal cord weakness/paresis s/p injection (2021), GERD" }
Output: { "topic_1": "Comorbidities", "topic_2": "Procedures/Surgery", "topic_3": "Disease Progression status" }
[..]

Figure 7: System prompt for question topic classification.

Thank you for helping us annotate our question dataset. Your time and effort are invaluable contributions to deploying robust LLM tools for clinical workflows.

You will be assessing the quality of LLM generated questions for a specific patient based on clinical relevancy, accuracy, and verifiability. These questions were generated in the context of admitting a patient to the hospital. More specifically, questions were generated based on information *prior* to a patient's most recent H&P note.

Dataset

When annotating each question, you will also have access to the following information:

MRN	Patient identifier to link to questions to Epic Hyperspace. **Make sure you review records that are records prior to the Reference H&P Timestamp.** Please do not share.
Reference H&P Timestamp	Date and time of the patient's most recent History & Physical (H&P) note. Records were sampled from a de-identified database so the timestamp of the real H&P Note in Epic Hyperspace may be a few days off.
Visit type	Type of visit. For some patients, there may be no visit type that was found.
Visit reason (unverified)	LLM extracted information from the H&P note. Use with caution.
Brief summary (unverified)	LLM extracted clinical summary from the H&P note. Use with caution.
Question Type	Reasoning required to answer the

Figure 8: Page 1 of 6: Annotation Guidelines

	question · Single hop (recent): question about a single, recent event · Single hop (past): question about a single, past event · Multi-hop: question requiring synthesis of multiple events / timepoints
Question Topic(s)	Identified topics. Multiple items may be chosen by the following: · Comorbidities · Procedures/Surgeries · Devices/Implants · Radiology/Imaging · Diagnostic testing (e.g. genetics, pathology, etc.) · Demographics ·Prescriptions (e.g. type, interactions, side effects, reasons not to medicate) · Laboratory tests · Disease Progression status (e.g. severity, complications, staging information, functional status) · Social Determinants of Health ·Assessment & Plan · Vitals · Appointments · Family History · Communications · Payment

Figure 9: Page 2 of 6: Annotation Guidelines

	· Reason for care (e.g. reason for admission, referral etc.) · Immunizations · Allergies
Answer (unverified)	The generated answer to the question. While we do not ask you to assess the answer, it may inform question assessment.
Supporting facts (unverified)	List of patient facts from the patient record that was used to generate the question. These facts have not been clinically verified.

As you review each question, we encourage you to look up the patient in Epic Hyperspace using the provided MRN to get the full clinical context. We hope this additional information will aid you during assessment.

Please filter for records prior to the Reference HP Timestamp to prevent data leakage.

Evaluation

For each question, you will be asked to answer a few binary questions. We have added some more information about each criteria below, in addition to positive and negative examples.

Question	Things to look for	Positive Example	Negative Example
Is this question clinically relevant for this patient	· bad questions cover topics that are not relevant to the current	What medications has this patient tried for her pelvic and rectal pain in	What were the findings of the patient's chest xray

Figure 10: Page 3 of 6: Annotation Guidelines

during the reference visit?	admission (i.e. wouldn't plausibly inform care) · good questions can be asked by *any* specialty that would reasonably see this patient during present admission	the last 3 years? [Patient was admitted with lower abdominal pain]	in 2021? [Patient was admitted with sudden left arm weakness and slurred speech]
Are all events referenced in the question consistent with the chart?	· references to events or diagnoses that are out of place · hallucinated information	What were this patient's right ventricular pressures following right pulmonary artery stenting in 2017?	What medications have been prescribed for the patient's heart disease? [Patient is 25M admitted after car crash, no history of heart disease]
Does the question have a well defined answer?	· multiple acceptable answers · imprecise language · unclear timeframes / time points	What was the working diagnosis for this patient's vertigo in 2015?	What complications did the patient experience following her cesarean delivery? [Complications is a vague term and it is unclear which delivery the question is referring to. Consider rephrasing to "Did this patient experience a postpartum hemorrhage, defined by

Figure 11: Page 4 of 6: Annotation Guidelines

			QBL/EBL >1000?"]
Is the question correctly classified as requiring single-hop (recent or past) or multi-hop reasoning?	· multi-hop questions that refer to events that happen in the same visit · single hop questions that refer to multiple events / time stamps		· Single hop (recent): What was the patient's recorded weight in 2021? [Patient admitted in 2022, question assumed most recent weight measurement was in 2021] · Single hop (past): What is the working diagnosis for this patient's chronic joint pain? [Patient could have multiple recorded diagnoses at different timepoints] · Multi-hop: What was prescribed following the patient's tonsillectomy in 2018?
Is the question correctly categorized by information type (e.g., labs, vitals, imaging)?	· select up to three topics · focus on the primary topics in the question		

Figure 12: Page 5 of 6: Annotation Guidelines

Could this question be rephrased to be more realistic, without changing the semantic meaning?	· Correct for awkward phrasing (times, over specification, etc.)		
Should this question be included in the benchmark?	· look for questions that are clinically relevant for the patient (at any possible visit), verifiable, and consistent with the chart.		

For each of the questions, if you answer "No," we will ask you to give an explanation or an alternative phrasing. **These free text explanations are strictly optional. If answering these free text explanations is too time consuming, do not feel the need to complete all of them.**

We have also added some optional questions for answer verification. These questions are optional:

- Is the information in the answer found in the chart?
- Is there missing information in the answer?

Figure 13: Page 6 of 6: Annotation Guidelines

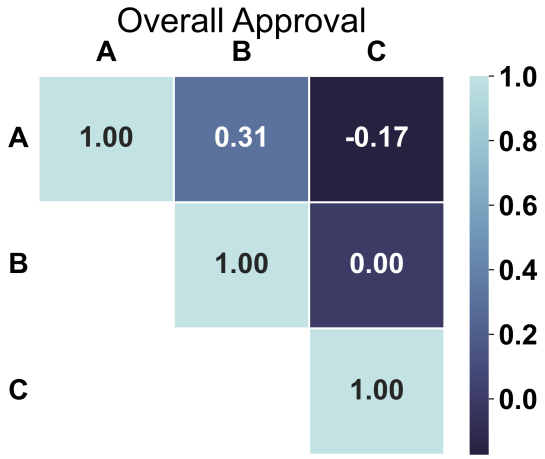

Figure 14: Cohen's κ for Annotator Responses for Overall Approval

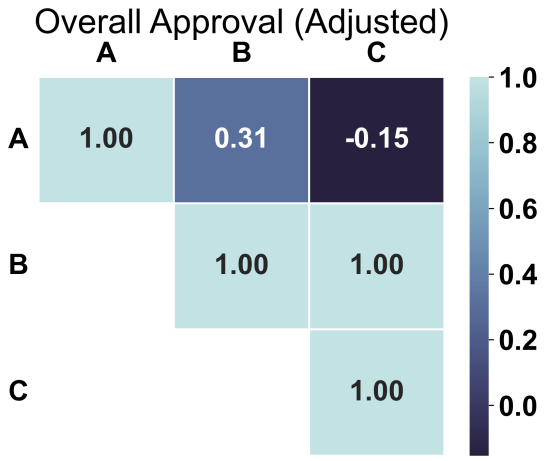

Figure 15: Cohen's κ for Annotator Responses for Overall Approval (Adjusted)

annotators, with a slight agreement of $\bar{\kappa} = 0.13$ (Figure 19). We may improve this agreement by using the rephrased questions collected during annotations as in-context examples (Section 4.3).

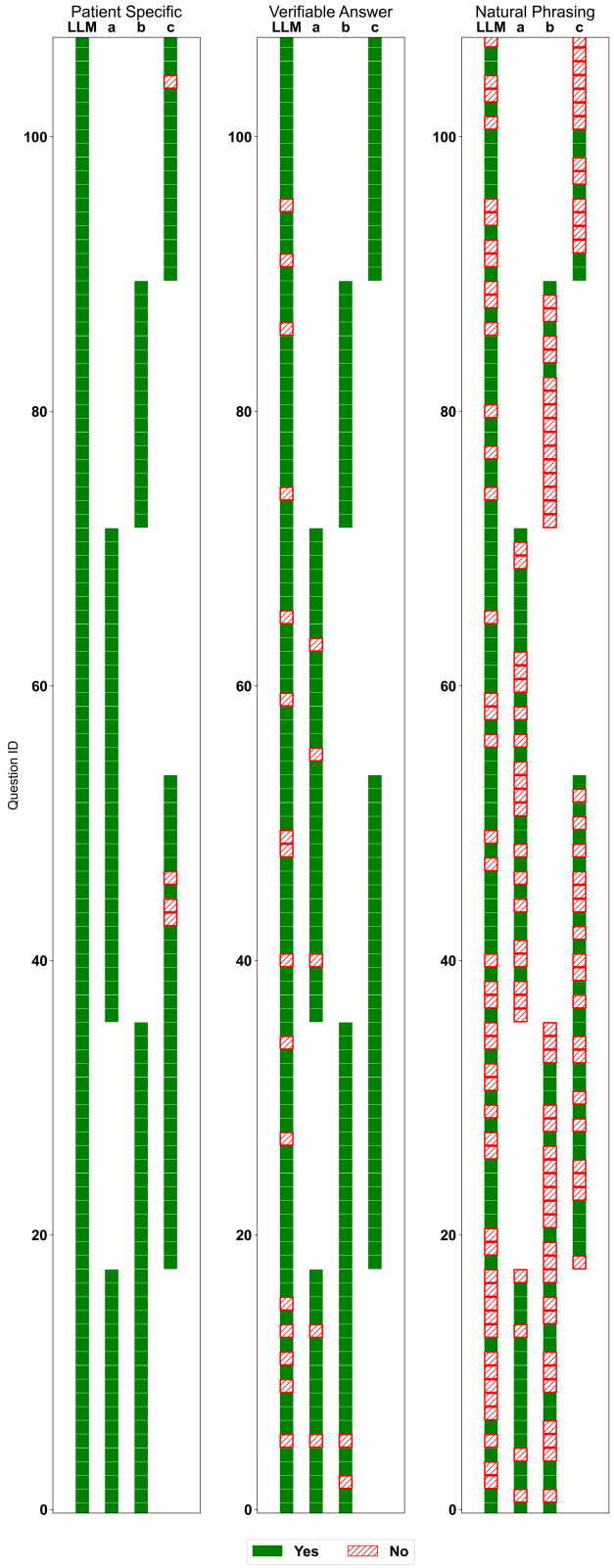

Figure 16: Annotator responses for the following assessment metrics: Patient Specific (left), Verifiable Answer (center), Natural Phrasing (right)

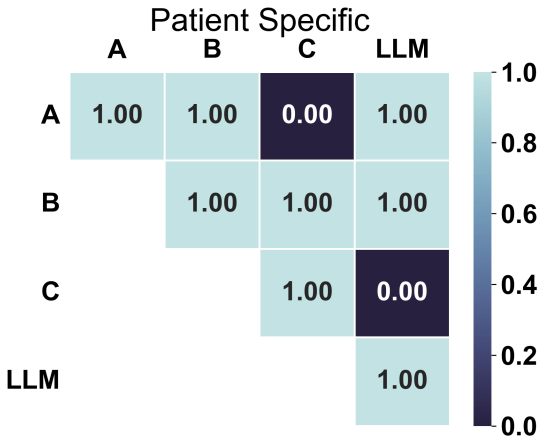

Figure 17: Cohen's κ for Annotator Responses for Patient Specific Questions

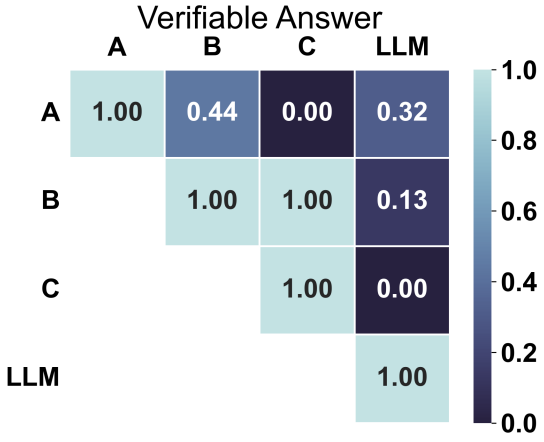

Figure 18: Cohen's κ for Annotator Responses for Verifiable Answers

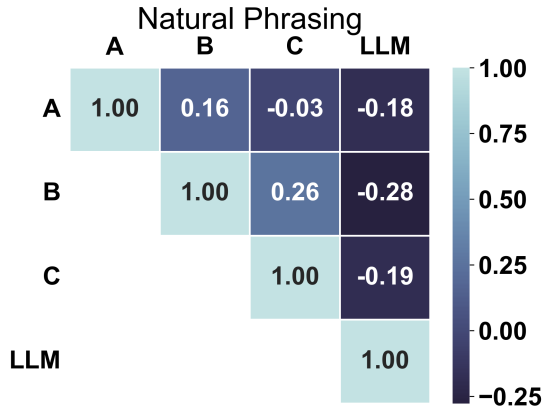

Figure 19: Cohen's κ for Annotator Responses for Natural Phrasing

LLM-as-a-Judge Prompt

You are a clinician reviewing automatically generated questions for an EHR retrieval benchmark. Your task is to decide if each question meets three criteria:
Clinical Relevance – Would a clinician reasonably want this information at the specific visit?
-"Yes" → The question is directly relevant to the reason for visit or note at this visit. If the question is ONLY relevant for the clinical summary, do NOT answer "Yes." If the information is mentioned in the note but does not pertain to the patient's immediate concerns, do NOT answer "Yes"
-"No, but for another visit" → The information is relevant to the patient's care at another point in time but not at this visit. This includes questions that pertain only to the clinical summary.
-"No, never relevant" → The question is clinically irrelevant to any visit.
Question Definition – Is the question clear, specific, and unambiguous?
-"Yes" → The question is narrowly scoped and can be linked to an objectively correct piece of information. Different clinicians would give the same answer.
-"No" → The question is too vague, underspecified, or too broad (e.g., "tell me about…", "details of…"). This question uses terminology that is subjective and not tied to a specific diagnosis or event (i.e. "interventions", "details", "medical history", "complications", "key factors" etc.). If multiple valid answers could exist, reject.
Natural phrasing – Is the question asked in the way a clinician would ask the question?
-"Yes" → The question is concise, uses clinical terminology, and mirrors how a clinician would actually search or ask (e.g., "What was the blood pressure at this visit?", "When was the last colonoscopy?").
-"No" → The question is awkward, verbose, or phrased in a way a clinician would not naturally use (e.g., "Provide details about the patient's hypertension management over time", "Summarize important factors about the case").
Decision Rules
-If unsure, choose the stricter option ("No, but for another visit" or "No").
-Do not reward questions that are vague, overly broad, or disconnected from the visit context.
-Explanations must cite which parts of the input (reason for visit, summary, note, or facts) support your decision.
Input Format "question": "string – Question (the LLM-generated question)",
"answer": "string– Example Answer (the potential response extracted from the EHR)",
"reference_timestamp": "string, ISO 8601 – Timestamp (date/time of the visit)",
"reason_for_admission": "string – Reason for visit",
"clinical_summary": "string – Patient clinical summary",
"visit_type": "string – Visit type (e.g., outpatient, inpatient, ED)",
"note": "string – H&P note (History & Physical for the visit)"

Output Format
Return a single JSON object with the following keys:
"question-relevance": "Yes — No, but for another visit — No, never relevant", "question-defined": "Yes — No", "question-rephrase": "Yes — No", "explanation": "Short justification for both decisions"
Example 1: [...]

Figure 20: System prompt for the LLM-as-a-Judge Module. Examples redacted to prevent PHI leakage.